DOCUMENT RESUME

ED 126 124 TM 005 379

AUTHOR Flanagan, John C.
TITLE Changes in School Levels of Achievement: Project TALENT Ten and Fifteen Year Retests.
PUB DATE [Apr 76]
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Academic Achievement; Achievement Tests; *Comparative Analysis; *Longitudinal Studies; Research Problems; *Sampling; Secondary Education; *Standardized Tests; Statistical Analysis; *Student Testing; Testing Problems; *Test Results; Time; Trend Analysis
IDENTIFIERS *Project TALENT; Test Score Decline

ABSTRACT

Project TALENT, a large scale, long range, longitudinal study, is designed primarily to provide insights into the characteristics and development of American adolescents. In order to compare trends in test results, a comprehensive battery of psychological, educational, and personality measures were administered to participants in the 1960, 1970, and 1975 studies. Specifically, the tests measured vocabulary, English, reading comprehension, creativity, mechanical reasoning, visualization, abstract reasoning, quantitative reasoning, mathematics, and computation. Between 1960 and 1975, performance on all language tests--vocabulary, English, and reading comprehension--dropped significantly. Reading comprehension mean scores dropped one or two percentile points. Scores in computation and quantitative reasoning also showed substantial declines. The abstract reasoning test showed gains of eleven and eight percentile points for males and females, respectively. Similarly, the creativity test showed gains for both sexes. The results of the study as well as its implications for further research are discussed. (BC)

22 April 1976

CHANGES IN SCHOOL LEVELS OF ACHIEVEMENT: PROJECT TALENT TEN AND FIFTEEN YEAR RETESTS

John C. Flanagan
American Institutes for Research
Palo Alto, California

The field of education has been reliving in recent months some of the experiences of those attempting to predict election results in the 1930's. The Literary Digest sent mail ballots to lists of several million automobile owners and telephone subscribers. They published the returns from these mailings each week and for a number of elections proudly pointed to the accuracy of their poll based on millions of straw-ballots. Then came the Roosevelt-Landon presidential election of 1936 and their poll involving more than 2 million voters was found to be ridiculously in error. On the other hand in this election George Gallup established the point that with a scientifically drawn and weighted sample of only about 1500 voters, it was possible to predict correctly the outcome of the election. The present paper reports on the results from two sampling studies comparing results from relatively small 1970 and 1975 samples with those obtained from the national Project TALENT sample in 1960.

These studies may be contrasted with the Literary Digest type that are now receiving a lot of attention in the press which interpret college admission test data based on voluntary applicant groups as if they were representative samples of the population. Further, the discussion in the press fails to identify specifically the types of abilities involved in these tests. To draw inferences about a population, a random or representative sample of that population is needed, not merely a large sample. Similarly, it is of only limited interest to report that high school students in 1975 know less about oranges than the students of 1960 knew about apples.

Another factor complicating many of the comparisons is the problem of equating or getting scores on a common scale when similar but not identical tests are used. The problems of equating and scaling and the need for systematic control have been comprehensively treated in previous

Presented at the Symposium of Division D and NCME during the 1976 meetings of the American Educational Research Association (AERA) in San Francisco, April 22, 1976

contributions.[1] These discussions can be summarized by noting that if the contents are identical in factorial composition, the character of the equating groups doesn't matter. When this is not the case the scaling or equating can be generalized only to groups having precisely the same competencies.

These considerations make it a relatively hopeless exercise to try to make any useful generalizations about education from data obtained on large samples of self-selected individuals. To get a measurement of the trend in one important educational outcome we asked 20 percent of the group of about 1000 schools that participated in Project TALENT in March 1960 to give us one hour's time of each of their eleventh grade students. Thus, about 13,000 students took the Project TALENT Reading Comprehension Test in March 1970 and answered some of the same questions on their educational experiences as did the students in these schools 10 years earlier. The mean scores from the schools were weighted to represent the complete population of all the students in the more than 20,000 secondary schools in the country.[2]

The results can be summarized by saying that there was a very slight gain in Reading Comprehension over this period. This was 0.5 raw score points (on a 48-item test) for the male eleventh grade group and 0.3 raw score points for the females. Three of the reports from students seem especially relevant to the present discussion. (1) The students in 1970 said that they studied a little less than the students in 1960. In 1970, 66 percent of the boys and 53 percent of the girls reported that "on the average I study less than 10 hours a week including study periods in schools as well as studying at home." This figure was 6 percentage points larger for boys and 4 percentage points larger for girls than the corresponding figures for 1960. (2) They said that they read about as many books as the 1960 group. In both 1960 and 1970, about 28 percent of boys and girls reported that they had read 11 or more books in the past 12 months. About 12 percent in 1960 and 10 percent in 1970 indicated that they had read no books in the preceding 12 months. (3) The students in 1970 reported slightly better attendance than did the students in 1960. Only about 10 percent said that they were absent 15 or more days in the last school

---

1 Flanagan, John C., "Units, Scores, and Norms" in Educational Measurement, E. F. Lindquist, editor, American Council on Education, Washington, D. C., 1951; and Flanagan, John C., "Obtaining useful comparable scores for non-parallel tests and test batteries" in Journal of Educational Measurement, 1964, 1, (1), 1-4.

2 Flanagan, John C. & Jung, S. M., Progress in education: A sample survey (1960-1970), Palo Alto, California: American Institutes for Research, 1971.

year as compared to 12 percent in 1960. Thus the 1960-1970 comparisons showed only slight changes in reading comprehension, study habits, reading habits, and school attendance.

An opportunity arose to retest students in a relatively small number of Project TALENT schools in the spring of 1975 as part of the standardization of a new guidance program being developed by the American Institutes for Research that used the 11-year follow-up data from Project TALENT. The names of the 17 schools and the grades tested are shown in Figure 1. About 1800 students in Grades 9, 10, and 11 in schools that had participated in Project TALENT in 1960 were given 10 of the same ability tests that were administered in these schools in 1960. The use of the same schools in 1960 and 1975 eliminates school sampling errors, since the two samples are identical. There remains the question of possible trends in the socio-economic level of the communities served by these schools. Small adjustments to the grade means were made in accordance with the changes reported by the principals. The effects of these were negligible in most cases. After these adjustments were made it was possible to make comparisons of the 1960 and 1975 scores for the 10 tests in three grades (9, 10, and 11) for each sex group. This yielded 60 comparisons.

To obtain an estimate of the extent to which the findings from this set of schools could be relied on to represent the results that would be obtained if all schools were retested in 1975, the schools were randomly divided into two comparable halves. The mean scores made by the students in each grade in each school for males and for females in the 10 ability measures were weighted by the number participating in that sub-group in 1975. The same weight was applied to the mean for each sex in each grade for the school in 1960 and 1975 to make the results comparable. The two independent estimates of the amount and direction of change for each test in each grade were divided by the appropriate standard deviation for the grade and sex to get these changes in comparable standard score units. This gave us six pairs of independent estimates, one for males and one for females in each of the three common grades 9, 10, and 11. Since there are 10 ability measures from Project TALENT being compared, this produces 60 pairs of independent estimates of change. The correlation coefficient between these 60 pairs was calculated.

This coefficient can be regarded as an estimate of the reliability coefficient for the differences obtained from these samples including half the

Figure 1

Project TALENT Schools Participating in 1975 Retest and Standardization Study

Schools	Location	Grades Tested
Shaw Jr. High	Swampscott, Mass.	(9)
Johnson City High	Johnson City, N. Y.	(9, 10)
West Greene High	Rogersville, Pa.	(9, 10, 11)
Edison Jr. High	East Gary, Indiana	(9)
Nowata Jr. High	Nowata, Oklahoma	(9)
Lincoln Jr. High	Orem, Utah	(9)
Horace Mann Jr. High	W. Allis, Wisconsin	(9)
Gorham High	Gorham, Maine	(10)
Greene High	Greene, N. Y.	(10, 11)
Huron High	Huron, Ohio	(10, 11)
Beaver Falls High	Beaver Falls, Pa.	(10, 11)
Monaca High	Monaca, Pa.	(10, 11)
Dublin High	Dublin, Ga.	(10, 11)
Hondo High	Hondo, Texas	(10, 11)
Central Valley High	Central Valley, Ca.	(10, 11)
Puyallup High	Puyallup, Washington	(10, 11)
Hamilton High	Trenton, New Jersey	(11)

schools. The coefficient is 0.81. Corrected by the Spearman-Brown formula to represent the reliability of an estimate based on both halves, this coefficient becomes 0.90. Using Formula 11:19 in Truman Lee Kelley's Fundamentals of Statistics, Cambridge, 1947, the best estimate of the true difference between the 1960 and 1975 scores is the product of this reliability coefficient, 0.90, and the observed difference when the latter is expressed as a deviation from the mean of all of the observed differences. It should also be noted that the correlation between the estimates of the changes obtained from the sample of schools and the true values of the changes that would be obtained if all the schools in the country were included in the study is the square root of this value which is 0.95. The substantial agreement found between the two series of independent estimates of change strongly supports the validity of these estimates of the changes between 1960 and 1975.

The results of the comparisons are shown in Table 1. Included in these changes is the very small adjustment to allow for the trends in the quality of the population in the community served by the schools as estimated by their principals. For the 17 secondary schools, four principals estimated that the quality of the community was "slightly lower," 10 that it was about the same, two that it was "slightly higher," and one that it was "higher" in 1975 than in 1960. The principals were asked to report changes of 5 to 10 percentile points as "slightly higher" or "slightly lower" and 10 percentile points or more as "higher" or "lower." Over all, this indicated little change. The weighted average changes were converted to raw score changes nearly all of which were less than one tenth of a point. These changes in quality of community were excluded from the apparent school trends before comparing the 1975 results with the 1960 results for these schools.

To make them more meaningful, these raw score changes have been converted to standard score units and to percentiles, corresponding to differences in the percentile values of the two mean scores. It is clear that between 1960 and 1975, performance on all the language tests -- Vocabulary, English, and Reading Comprehension -- dropped significantly. The largest drops are in Vocabulary and English; they are about 0.4 and 0.3 in terms of standard scores for the two sexes combined. The corresponding mean scores are between 11 and 17 percentile points lower in 1975 than in 1960 in terms of the national percentile norms for the 10th grade. In contrast, the Reading Comprehension mean scores

have dropped about half a point, or about .05 standard score units, or 1 to 2 percentile points.

Other tests that showed substantial declines between 1960 and 1975 were Computation and Quantitative Reasoning. These declines ranged from 17 percentile points for the males on Computation to 8 percentile points for both males and females on the Quantitative Reasoning Test. It is interesting to note that in contrast to these results the Mathematics test showed a small gain of 2 to 3 percentile points rather than a decline for this period.

Table 1

The Comparisons of 1960 and 1975 Results in Grades 9, 10, and 11 of Students in 17 Secondary Schools Using the Same Project TALENT Tests

	Males (N_{1975} = 871)					Females (N_{1975} = 926)				
	Raw Score		Change 1975-1960	Standard Score	%ile Difference	Raw Score		Change 1975-1960	Standard Score	%ile Difference
	1960	1975				1960	1975			
Vocabulary	18.5	15.7	-2.8	-.48	-17%	17.3	15.5	-1.8	-.32	-11%
English	77.3	73.4	-3.9	-.26	-12%	84.5	79.7	-4.8	-.34	-16%
Reading Comprehension	28.8	28.4	-0.4	-.04	- 1%	29.8	29.1	-0.7	-.07	- 2%
Creativity	9.1	10.0	0.9	.22	8%	8.4	10.1	1.7	.46	16%
Mechanical Reasoning	12.4	12.2	-0.2	-.05	- 2%	8.5	9.2	0.7	.20	7%
Visualization	8.9	8.8	-0.1	-.03	- 1%	7.8	8.1	0.3	.10	4%
Abstract Reasoning	8.7	9.5	0.8	.26	11%	8.7	9.4	0.7	.22	8%
Quantitative Reasoning	8.5	7.8	-0.7	-.20	- 8%	8.0	7.2	-0.8	-.23	- 8%
Mathematics	10.5	10.7	0.2	.04	2%	9.9	10.3	0.4	.10	3%
Computation	25.7	18.7	-7.0	-.27	-17%	30.8	26.9	-3.9	-.19	-11%

The Abstract Reasoning test showed gains of 11 and 8 percentile points for the males abd females, respectively. This is a non-verbal test using the progressive matrices principle. Similarly, the Creativity test shows gains for both sexes. In this case the females gained 16 percentile points and the males gained 8 percentile points.

Both of the remaining two tests have been found predictive of success in mechanical work. Interestingly, both show modest gains for the females and very slight losses for the males. The females gained 7 and 4 percentile points respectively in Mechanical Reasoning and Visualization and the males lost 2 and 1 percentile points on these tests. This appears to reflect a change away from the sex stereotyping that characterized the 1960 group. It should be noted that the males still have developed these abilities to a much greater extent than the females, these changes reduced the differences favoring the males by only about one-fourth and one-third respectively for these two tests.

This is clearly not regarded as a definitive study on changes in abilities during recent years. It might better be regarded as a pilot study or prototype of the needed types of data collection. One finding that points up the need for precise definition of the abilities being compared is the disparate results for the various types of language and quantitative tests. To provide further evidence on the nature of these variables a fairly typical set of correlations based on about 950 10th grade students is shown in Table 2. The amount of overlapping and the amount of unique variance in each of these types of measures as indicated by the intercorrelations in Table 2 are more precisely shown in Table 3.

Experience with these measures over the past 15 years indicates that the tests of English and Computation represent the students' performance on a fairly straightforward and simple type of learning. Thus, these two tests, to a greater extent than the others, show how conscientious the student has been in doing his or her assignments. They reflect especially study time on these tasks and the obvious conclusion is that this has been reduced over the past 15 years. The decline in Vocabulary suggests less time reading both in and out of school. The 1970 survey indicated less study time, but did not show a reduction in outside reading.

The relative stability over this period of the proficiency of the students in reading comprehension and mathematics is of greatest importance since what

Table 2

Intercorrelation Coefficients for the 10 PCG Ability Measures (Grades 8-11)

Grade 10

	Vocabulary	English	Reading Comprehension	Creativity	Mechanical Reasoning	Visualization	Abstract Reasoning	Quantitative Reasoning	Mathematics	Computation
Vocabulary	1.000	0.684	0.793	0.620	0.578	0.496	0.560	0.618	0.645	0.314
English	0.684	1.000	0.671	0.482	0.387	0.375	0.481	0.550	0.588	0.445
Reading Comprehension	0.793	0.671	1.000	0.649	0.562	0.503	0.577	0.622	0.632	0.334
Creativity	0.620	0.482	0.649	1.000	0.589	0.531	0.536	0.544	0.541	0.279
Mechanical Reasoning	0.578	0.387	0.562	0.589	1.000	0.534	0.566	0.511	0.518	0.180
Visualization	0.496	0.375	0.503	0.531	0.534	1.000	0.627	0.512	0.501	0.264
Abstract Reasoning	0.560	0.481	0.577	0.536	0.566	0.627	1.000	0.538	0.624	0.306
Quantitative Reasoning	0.618	0.550	0.622	0.544	0.511	0.512	0.538	1.000	0.671	0.391
Mathematics	0.645	0.588	0.632	0.541	0.518	0.501	0.624	0.671	1.000	0.507
Computation	0.314	0.445	0.334	0.279	0.180	0.264	0.306	0.391	0.507	1.000
Mean	39.744	31.369	25.266	11.258	12.721	15.260	12.761	11.667	12.802	29.041
Standard Deviation	10.983	6.997	8.352	4.162	4.323	4.759	4.213	3.674	4.998	13.742
Number of Cases	954	957	962	956	956	953	955	957	957	955

Table 3

Analysis of the Proportion of the Unique Variance of Each Ability Test and the Potential Unique Validity Coefficient of Each Test with a Pure Criterion Measure of Its Unique Function

TENTH GRADE (N= 957)

	Error Variance $1 - r_{iI}$	Multiple Correlation (9 tests) R_{ic}	Overlapping Variance $\frac{R^2_{ic}}{r_{cC}}$	Unique Variance $r_{iI} - \frac{R^2_{ic}}{r_{cC}}$	Potential Unique Validity * $\left(r_{iI} - \frac{R^2_{ic}}{r_{cC}}\right)$
Vocabulary	.07	.84	.75	.18	.42
English	.14	.75	.59	.27	.52
Reading Comprehension	.13	.84	.73	.14	.37
Creativity	.27	.73	.56	.17	.41
Mechanical Reasoning	.29	.71	.55	.16	.40
Visualization	.17	.69	.53	.30	.55
Abstract Reasoning	.23	.75	.60	.17	.41
Quantitative Reasoning	.25	.74	.57	.18	.42
Mathematics	.14	.80	.68	.18	.42
Computation	.17	.57	.38	.45	.67

* Maximum possible validity coefficient of test with pure criterion measure of test's unique function.

these tests measure are probably the two most important abilities for effectiveness in adult roles. The fairly substantial increases shown with respect to abstract reasoning and creativity are intriguing. Without further data one can only speculate that perhaps some of the attention being given by teachers and students to problem solving and project activities of a more practical nature is beginning to show results. The decline in quantitative reasoning could be related to the changes the new math has introduced with respect to the type of "reasoning" problems being studied. The other changes were discussed earlier.

What conclusions can be drawn from this prototype study? Clearly it is essential to study trends in all of the abilities being developed by the present generation of students and relate these to all their educational experiences. The program of the National Assessment of Educational Progress is designed to tell us of trends in educational outcomes. However, it tells us nothing about what is causing these changes. It is essential to formulate and test specific hypotheses as to why these changes are taking place. Vocabulary has been shown to be largely a function of amount and type of reading both in school and out. What are people reading? How much is television reducing reading time? Are students in fact spending less time on learning to add, subtract, multiply, and divide? What are they doing instead of learning the rules of punctuation, capitalization, and language usage? What specific activities are resulting in the women's gains in creativity, mechanical reasoning, and visualization? What can we learn about the whole field of problem solving, reasoning, and mathematics from relating specific student activities with specific educational outcomes? Are the newer courses adding valuable insights or just taking away time from the basic skills?

It is essential to have intensive studies of individuals on a longitudinal basis, in order to relate changes in student performance to specific student experiences and provide a sound basis for ensuring progress in education. The prototypes for which some results have been presented here should be applied in systematic and comprehensive studies. As the consistency of the foregoing results show, the numbers do not have to be very large. However, if the effects of varying types of educational and out-of-school activities are to be accurately measured they cannot be nearly as small as in the 1975 study.

APPENDIX

Descriptions of the Ten Ability Measures

Vocabulary

The 60 items of the Vocabulary section measure the individual's knowledge of meanings of words. The items require the examinee to identify synonyms, select examples of stimulus words, or recognize operational definitions.

132. Obstruction means
- A. obeisance.
- B. obstacle.
- C. obstinacy.
- D. obstreperousness.
- E. obtuseness.

English

Spelling, capitalization, punctuation, English usage, grammatical correctness, and clarity of expression are measured in the 48 items of the English section. The format is that of a blank in a stimulus sentence with five choices offered as options for filling that blank. Standard English usage is the basis for selecting the right answer.

14. Neither of them _____ there yet.
- A. had went
- B. have gone
- C. has went
- D. have went
- E. has gone

Mathematics

The 24 items of the Mathematics section assess knowledge of and ability to apply elementary concepts in mathematics.

24. If s is the side of a certain square, the side of a square whose area is four times as large equals
- A. $s + 2$
- B. $2s$
- C. $s + 4$
- D. $4s$
- E. s^2

Reading Comprehension

The Reading Comprehension section contains 40 items that measure literal and critical comprehension of passages. The stimulus material was selected with the criteria that the material be unfamiliar to most high school students and that it pertain to interesting subject areas.

(1) Literature is the art of selection and ruthless ex-
(2) clusion, or, as Flaubert has said, it is "the art
(3) of making sacrifices." Chekhov, another enemy of
(4) the unnecessary word, says: "If in the first chapter
(5) you say that a gun hung on the wall, in the second
(6) or third chapter it must without fail be discharged."
(7) He is right. The writer who seeks to emulate de
(8) Maupassant and Poe in producing stories with pre-
(9) cisely controlled plots cannot afford to ignore his
(10) advice.

38. Who makes the sacrifices referred to in line 3?
- A. Literary critics
- B. Bad writers
- C. Good writers
- D. Bad readers
- E. Good readers

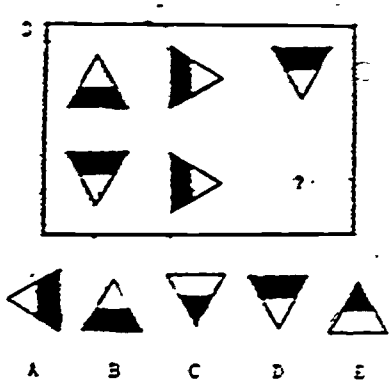
Quantitative Reasoning

The Quantitative Reasoning section contains 22 items that measure the individual's ability to solve word problems in mathematics. In most problems it is necessary to identify the method of solution rather than the actual numerical answer.

8. A salesman gets 12% commission on all items sold over a quota of 5. If he sells 8 items at $30 each, his commission is 12% of
- A. $240
- B. $90
- C. $30
- D. $8
- E. $5

Abstract Reasoning

The Abstract Reasoning section contains 22 items that measure one's ability to detect sequences and patterns. The stimuli are sets of abstract geometric figures, with one or more figures missing.



Computation

The Computation section, a highly speeded test, measures an individual's accuracy and speed in addition, subtraction, multiplication, and division. To determine the raw score, three points for every wrong answer are subtracted from the number right.

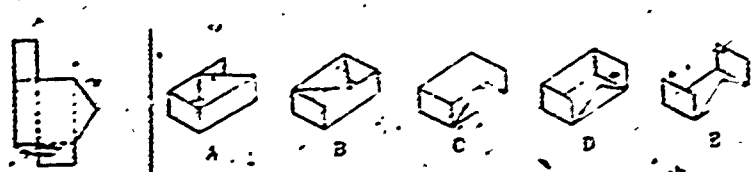
43. Multiply:

$$\begin{array}{r} 68 \\ \times\ 14 \\ \hline \end{array}$$

A. 796
B. 832
C. 852
D. 876
E. 952

Visualization

The Visualization section has 25 items that measure an individual's ability to visualize in three-dimensional space. Each item has a drawing of a flat shape and of five three-dimensional objects. The task is to select the object that could be formed by folding the shape along dotted lines or rolling the shape or both.



Creativity

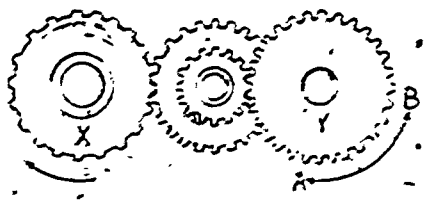
The 24 items of the Creativity section require the individual to think of innovative ways of accomplishing tasks or of new uses for simple objects. A problem situation is presented along with answer choices that have the first and last letters given and each intervening letter designated by a blank. The individual must think of a solution that fits into one of the choices offered.

11. Intricate systems of circuits and wiring often need quick repair in emergencies. Part of the maintenance problem has been overcome by making it possible to spot the area where trouble occurs quickly and repair that section. Since the identification of such small wires by written labels is impossible, they are identified by using

A. s - - - - s
B. d - - - - s
C. c - - - - s
D. l - - - - s
E. m - - - - s

Mechanical Reasoning

The Mechanical Reasoning section contains 25 items that assess the ability to understand mechanical ideas by looking at pictures or diagrams. The stimuli are drawings of simple mechanical objects such as an olive press and a windmill.



8. When gear X turns in the direction shown by the arrow, gear Y turns

A. in direction A.
B. in direction B.
C. first in one direction and then in the other.